

Метрики моделей бинарной классификации

Матрица ошибок	Модель говорит: ноль (отрицательное)	Модель говорит: единица (положительное)	"Эпидемические" метрики: как точно модель отражает реальный мир?
	ошибки	ошибки	
На самом деле: ноль (отрицательное значение)	TN	FP I рода	→ Specificity (специфичность), <i>Recall</i> нулевого класса Доля верно определенных отрицательных в выборке
На самом деле: единица (положительное значение)	ошибки FN II рода	TP	→ Recall (полнота), Sensitivity (чувствительность), True Positive Rate Доля верно определенных положительных в выборке, снижается с ошибками II рода (FN)
По смыслу: "ответственность модели" Доля на самом деле положительных среди предсказанных положительно, снижается с ошибками I рода (FP)		Precision	Будьте внимательны: порядок на осях как в sklearn, а не как в Википедии!
Подсказка к расчету метрик: <i>True-показатели (серый фон) — в числителе,</i> <i>сумма показателей на линии — в знаменателе</i>			

Дисбаланс классов. Примеры.

1. Дисбаланс классов 90/10, модель в 99% случаев предсказывает нулевой класс

	TN: 90	FP: 0	90	отрицательных в выборке
	FN: 9	TP: 1	10	положительных в выборке
	99	1	100	
	модель считает отрицательными	модель считает положительными		
Accuracy	$\frac{TN+TP}{Total}$	0.910	0.090	Misclassification rate = (1 - Accuracy)
Recall	$\frac{TP}{TP+FN}$	0.100	0.550	Balanced Accuracy = (Recall + Specificity) / 2
Specificity	$\frac{TN}{TN+FP}$	1.000	0.000	False Positive Rate = (1 - Specificity)
Precision	$\frac{TP}{TP+FP}$	1.000		
F1 Score	$\frac{2 * Precision * Recall}{Precision + Recall}$	0.182		

2. Дисбаланс классов 90/10, в каждом классе верно предсказано 80%

	TN: 72	FP: 18	90	отрицательных в выборке
	FN: 2	TP: 8	10	положительных в выборке
	74	26	100	
	модель считает отрицательными	модель считает положительными		
Accuracy	$\frac{TN+TP}{Total}$	0.800	0.200	Misclassification rate
Recall	$\frac{TP}{TP+FN}$	0.800	0.800	Balanced Accuracy
Specificity	$\frac{TN}{TN+FP}$	0.800	0.200	False Positive Rate
Precision	$\frac{TP}{TP+FP}$	0.308		
F1 Score	$\frac{2 * Precision * Recall}{Precision + Recall}$	0.444		

Построение ROC-кривой

ROC расшифровывается (хотя, скорее зашифровывается) как Receiver Operating Characteristic.

Модели-классификаторы возвращают (predict_proba в sklearn) значения в диапазоне от 0 до 1, которые удобно считать вероятностью класса. Решение об отнесении измерения к единичному классу принимается в зависимости от выбранного порога threshold. Если $p > \text{threshold}$, то считаем, что предсказан единичный (положительный) класс.

В зависимости от порога меняются метрики модели. Для визуализации изменений используется ROC-кривая. Каждая точка кривой — это значения пары метрик модели для одного из порогов. Количество точек равно выбранному количеству порогов. Принято начинать кривую в точке (0, 0) и заканчивать в точке (1, 1).

Метрика по оси X: False Positive Rate, по оси Y: True Positive Rate

Для построения кривой нужна таблица истинных классов, отсортированная по убыванию probability.

Метрика ROC_AUC вычисляется как площадь под ROC-кривой (area under curve). В идеале roc_auc = 1. Диагональная линия показывает, как в среднем работал бы классификатор на базе подбрасывания монетки.

Пример

Столбцы таблицы		Описание	
Yn		Если истинный класс положительный — 0, иначе — 1	
Yp		Если истинный класс положительный — 1, иначе — 0	
P		Значения probability, полученные от классификатора	
FPR		False Positive Rate, отношение текущего значения FP к максимальному в столбце	
FP		Число False Positive, рассчитанное в предположении, что порог строчкой ниже	
TP		Число True Positive, рассчитанное в предположении, что порог строчкой ниже	
TPR		True Positive Rate, отношение текущего значения TP к максимальному в столбце	
AUC		Часть площади под кривой между предыдущей и данной точкой	

Заданные значения				Рассчитанные значения				AUC	Предсказанный класс в случае threshold = 0.5
Yn	Yp	Истинный класс	P	FPR	FP	TP	TPR		
искусственная точка [0, 0]				0			0		
0	1	положительный	0.9	0	0	1	0.2	0	положительный
0	1	положительный	0.8	0	0	2	0.4	0	положительный
0	1	положительный	0.7	0	0	3	0.6	0	положительный
1	0	отрицательный	0.6	0.25	1	3	0.6	0.15	положительный
0	1	положительный	0.5	0.25	1	4	0.8	0	отрицательный
1	0	отрицательный	0.4	0.5	2	4	0.8	0.2	отрицательный
0	1	положительный	0.3	0.5	2	5	1	0	отрицательный
1	0	отрицательный	0.2	0.75	3	5	1	0.25	отрицательный
1	0	отрицательный	0.1	1	4	5	1	0.25	отрицательный

ROC_AUC = 0.85

